## Deliverable D3.1.1

**Early text annotation prototype**

<table>
<thead>
<tr>
<th>Editor:</th>
<th>Achim Rettinger, KIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s):</td>
<td>Achim Rettinger, KIT; Lei Zhang, KIT; Alexander Hageman, KIT; Marko Tadić, UZG; Željko Agić, UZG</td>
</tr>
<tr>
<td>Deliverable Nature:</td>
<td>Prototype (P)</td>
</tr>
<tr>
<td>Dissemination Level:</td>
<td>Public (PU)</td>
</tr>
<tr>
<td>(Confidentiality)¹</td>
<td></td>
</tr>
<tr>
<td>Contractual Delivery Date:</td>
<td>M12</td>
</tr>
<tr>
<td>Actual Delivery Date:</td>
<td>M12</td>
</tr>
<tr>
<td>Suggested Readers:</td>
<td>All partners using the XLike Toolkit</td>
</tr>
<tr>
<td>Version:</td>
<td>1.0</td>
</tr>
<tr>
<td>Keywords:</td>
<td>cross-lingual document linking, semantic annotation, conceptual space, Wikipedia, NER</td>
</tr>
</tbody>
</table>

¹ Please indicate the dissemination level using one of the following codes:
- **PU** = Public
- **PP** = Restricted to other programme participants (including the Commission Services)
- **RE** = Restricted to a group specified by the consortium (including the Commission Services)
- **CO** = Confidential, only for members of the consortium (including the Commission Services)
- **Restreint UE** = Classified with the classification level "Restreint UE" according to Commission Decision 2001/844 and amendments
- **Confidentiel UE** = Classified with the mention of the classification level "Confidentiel UE" according to Commission Decision 2001/844 and amendments
- **Secret UE** = Classified with the mention of the classification level "Secret UE" according to Commission Decision 2001/844 and amendments
Disclaimer

This document contains material, which is the copyright of certain XLike consortium parties, and may not be reproduced or copied without permission.

All XLike consortium parties have agreed to full publication of this document.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the XLike consortium as a whole, nor a certain party of the XLike consortium warrant that the information contained in this document is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using this information.

<table>
<thead>
<tr>
<th>Full Project Title:</th>
<th>XLike – Cross-lingual Knowledge Extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short Project Title:</td>
<td>XLike</td>
</tr>
<tr>
<td>Number and Title of Work package:</td>
<td>WP3 – Cross-lingual Semantic Annotation</td>
</tr>
<tr>
<td>Document Title:</td>
<td>D3.1.1 Early text annotation prototype</td>
</tr>
<tr>
<td>Editor (Name, Affiliation)</td>
<td>Achim Rettinger, KIT</td>
</tr>
<tr>
<td>Work package Leader (Name, affiliation)</td>
<td>Marko Tadić, UZG</td>
</tr>
<tr>
<td>Estimation of PM spent on the deliverable:</td>
<td>7 PM</td>
</tr>
</tbody>
</table>

Copyright notice

© 2012-2014 Participants in project XLike
Executive Summary

The main goal of the XLike project is to extract knowledge from multi-lingual text documents by annotating statements in sentences of a document with a cross-lingual knowledge base. The purpose of the early text annotation prototype described here, is to investigate the performance of shallow multi-lingual text annotation tools with a cross-lingual knowledge base, namely Wikipedia. The resulting baseline performance will be compared to the semantic annotation tool developed for D3.1.2. While this prototype does only annotate word phrases in the text documents and link them to Wikipedia pages in any language, the final annotation prototype will extract subject-predicate-object triples (output of D2.2.1 and D2.2.2) and link them to a semantic knowledge representation like Wikidata or Cyc. While this deliverable focuses on analysing the performance of existing multilingual approaches to cross-lingual annotation the functionality of the D3.1.2 prototype is more of an open research question.

In this document we investigate different tools based on the next 3 approaches for multi-lingual annotation:

1. Named Entity Recognition (NER): This approach is based on the Named Entities detected by D2.1.1. On top of that a simple approach for finding the corresponding Wikipedia pages in the target language is deployed.

2. Wikipedia Miner Wikifier (WIFI): This approach is trained on existing links in Wikipedia articles to detect similar phrases and links in any text document of the same language as the Wikipedia used for training. Again, a simple approach for finding the corresponding Wikipedia pages in the target language is deployed.

3. Cross-lingual Explicit Semantic Analysis (CLESA): While the first two approach detect word phrases this service links articles by topic to corresponding Wikipedia pages in the target language.

While all implemented tools support many languages (NER: all XLike languages; WIFI: English, German, Spanish; CLESA: all XLike languages) for practical reasons, the evaluation presented here, is focused on annotating German and Spanish documents only and linking them to the English Wikipedia. The English Wikipedia can be considered a hub knowledge base (but not a language independent knowledge base), as it is by far the largest and best linked (including language links) Wikipedia.

The goal of the evaluation was to measure the cross-lingual linking capabilities of the procedures developed. The basic assumption of evaluation process is that the state-of-the-art approaches for monolingual annotation and grounding in Wikipedia as conceptual space perform best on English documents and the English Wikipedia. Thus, the evaluation is focused on measuring the number of correct annotations from non-English documents (here German and Spanish) to the English Wikipedia, based on a parallel English-German-Spanish corpus, and comparing the found links to the ones found by the monolingual English annotation services.

From now on whenever we use the term “XLike languages” we refer to English, German, Spanish, Chinese, Catalan and Slovenian.
# Table of Contents

Executive Summary .................................................................................................................. 3  
Table of Contents .................................................................................................................... 4  
List of Figures and/or List of Tables ....................................................................................... 5  
Abbreviations .......................................................................................................................... 6  
Definitions ................................................................................................................................ 7  
1 Introduction .......................................................................................................................... 8  
   1.1 Motivation ....................................................................................................................... 8  
   1.2 STA use case .................................................................................................................. 8  
2 Techniques for cross-lingual shallow text annotation .......................................................... 9  
   2.1 General Framework ........................................................................................................ 9  
   2.2 Background data corpora ............................................................................................... 9  
   2.3 Named Entity Recognition based approach ................................................................. 9  
   2.4 Wikipedia Miner Wikifier based approach ................................................................. 10  
   2.5 Cross-lingual Explicit Semantic Analysis based approach ............................................ 10  
3 Cross-lingual Annotation Web Services .............................................................................. 11  
   3.1 Named Entity Annotation Service ................................................................................ 11  
   3.2 Wikipedia Miner Wikifier Annotation Service ............................................................ 12  
   3.3 Cross-lingual Explicit Semantic Analysis Service ....................................................... 13  
4 Evaluation of Cross-lingual Annotation ............................................................................ 15  
   4.1 Evaluation of Cross-lingual NER Annotation Service .................................................. 15  
   4.2 Evaluation of Cross-lingual WIFI Annotation Service ................................................. 17  
   4.3 Evaluation of CLESA Annotation Service .................................................................. 18  
   4.4 Summary of Evaluation ............................................................................................... 18  
References ................................................................................................................................ 20
List of Figures and/or List of Tables

Figure 1. Example output of the wikifier service................................................................. 10
Figure 2. Example input and output of the named entity annotation service.......................... 12
Figure 3. Example output of the cross-lingual ESA service .................................................. 14
Figure 4. Average number of NE links to English Wikipedia extracted per document ............. 16
Figure 5. Average precision of NE links to English Wikipedia............................................. 16
Figure 6. Average number of Wikifier links to English Wikipedia extracted from one document..... 17
Figure 7. Average precision of Wikifier links to English Wikipedia ..................................... 17
Figure 8. Average precision of CLESA links to English Wikipedia........................................ 18
Figure 9. Average combined precisions of all combinations of three approaches to cross-lingual linking to
   English Wikipedia articles ................................................................................................. 19

Table 1. Named Entity Annotation Service............................................................................... 11
Table 2. Wikipedia Miner Wikifier Annotation Service .......................................................... 13
# Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NER</td>
<td>Named Entity Recognition</td>
</tr>
<tr>
<td>WIFI</td>
<td>Wikipedia Miner Wikifier</td>
</tr>
<tr>
<td>CLESA</td>
<td>Cross-lingual explicit semantic analysis service</td>
</tr>
</tbody>
</table>
## Definitions

<table>
<thead>
<tr>
<th>Parallel Corpus</th>
<th>Parallel corpus consists of documents that are translated directly into different languages.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparable Corpus</td>
<td>Comparable corpus contains, unlike parallel corpora, no direct translations. Overall they may address the same topic but can differ significantly in length, detail and style.</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Motivation

The main goal of the XLike project is to extract formal knowledge from multi-lingual text documents by annotating statements in sentences of a document with a cross-lingual knowledge base. The purpose of the early text annotation prototype described here, is to investigate the performance of shallow multi-lingual text annotation tools with a cross-lingual knowledge base, namely Wikipedia. The resulting baseline performance will be compared to the semantic annotation tool developed for D3.1.2 in year 2. While this prototype does only annotate simple word phrases in the text documents and link them to Wikipedia pages in any language, the final annotation prototype will extract subject-predicate-object triples (output of D2.2.1 and D2.2.2) and link them to a semantic knowledge representation like Wikidata or Cyc. Such triples are essential to being able to apply logical constraints specified in a knowledge base or extract semantic graphs, e.g., using event patterns. However, here we do not relate a phrase to other elements in a sentence, but concentrate on recognizing and classifying single entities according to a cross-lingual knowledge base.

1.2 STA use case

STA covers topics related to Slovenia or Slovenian entities (E.g. companies, athletes). As such, tracking relevant news is an important part of editors’ daily routine. Technologies developed within XLike project can improve this process by providing tools for detecting relevant articles across languages and media (mainstream, social media).

Formally, topic or entity tracking can be seen as a filter applied to a stream of articles. An article is retained by the filter if it matches the topic, or is related to the entity. Topics can be defined as a standard classification task, with articles on the input and set of matching topics on the output. Entities can be detected using named-entity extractors.

For popular topics or entities, the filter can retain a large amount of articles. The information contained within these articles can be visualized or summarized to help the editors in skimming through the content, to identify relevant events.

Here, STA defined the list of entities to be tracked by a) links to Wikipedia in different languages or by b) textual descriptions of the entities to be tracked. Concerning case a) we can apply NER and WIFI to tackle the task while in case b) only CLESA can be used.
2 Techniques for cross-lingual shallow text annotation

2.1 General Framework

All techniques presented here rely at some point on a comparable corpus to establish the cross-lingual annotation. While the NER modules are first trained for each language individually on different corpora (see D2.1.1), in a second step the detected named entities are then grounded in a cross-lingual comparable corpus extracted from Wikipedia. In contrast, WIFI and CLESA are directly trained on Wikipedia and don’t need a second step that links entities to Wikipedia.

2.2 Background data corpora

For this early text annotation prototype we developed three approaches for cross-lingual document annotation. After we built up models and developed approaches, we executed some experiments on this three approaches and evaluated them. For the purpose of training the WIFI service we used comparable corpora from Wikipedia in different languages and for testing we used the parallel multilingual corpus JRC-Acquis.

Joint Research Centre Collection of the Acquis Communautaire, abbreviated as JRC-Acquis, is a multilingual parallel corpus extracted from Acquis Communautaire. Acquis Communautaire (AC) is a French term that means ‘the EU as it is’. This is a body of common rights and obligations which bind all the Member States together within the European Union as published in the Official Journal of the EU. The JRC-Acquis corpus is the biggest existing parallel corpus. The corpus is available in 23 languages, consists on average of more than 20,000 documents per language, with average size of nearly 46 million words per language. The corpus is available in TEI-compliant XML format, and consists of two parts, the marked-up texts and the bilingual alignment information for all the 190 language pairs. With the help of the EUROVOC thesaurus the JRC-corpus is classified manually into subject domains as they are defined in EUROVOC. The JRC-Acquis corpus is available for download at http://langtech.jrc.it/JRC-Acquis.html.

Wikipedia is currently the largest knowledge treasure on the web in the world and various editors develop it constantly, therefore its breadth and depth also are expanding continually. The documents at Wikipedia are available in 260 languages and they are linked to each other with cross-language links in case they describe the same topic. The most documents at Wikipedia are available in English there are currently more than 4 million pages. The German, Spanish, French, Italian and other 5 Wikipedias contain more than 750 thousand documents. The Chinese, Slovenian and other 18 Wikipedias contain more than 150 thousand documents. The Greek, Norwegian and other 8 Wikipedias have more than 50 thousand documents. The documents not only are available in large number of different languages also available in diverse domains. That is why we use Wikipedia corpus as the comparable corpus. The Wikipedia database dumps you can download at http://dumps.wikimedia.org/. There, the content, article-to-article link list, image metadata and misc bits are available for download in different XML wrapper formats. The download tool for Wikipedia database dumps can be found at https://github.com/babilen/wp-download/.

2.3 Named Entity Recognition based approach

This approach is based on the Named Entities detected by the NERC tools described in D2.1.1 for all XLike languages. On top of that a straight-forward approach for finding the corresponding Wikipedia pages in the target language is deployed. First the named entity string is used for a keyword search for the Wikipedia article in the same language and then the language links of the top-ranked Wikipage are used to link to the target language. Named Entities are used here as the current computationally inexpensive representatives of other types of entities that are going to be cross-lingually linked in the further development of this system.
2.4 Wikipedia Miner Wikifier based approach

There is a large body of work that tries to learn links from unstructured text to a collection of relevant Wikipedia topics given the Wikipedia as a training corpus (see [Mihalcea 2007]). One of the most prominent approaches is the one proposed by [Milne2008] achieving around .75 in precision and recall. Our cross-lingual annotation services are based on this method.

Karlsruhe is a city in the state of Baden-Württemberg, in southwest Germany, near the Franco-German border. Karlsruhe was founded in 1715 as Karlsruhe Palace, when Germany was a series of principalities and city-states.

## Figure 1. Example output of the wikifier service

There are two separate stages involved: link disambiguation and link detection. In the link disambiguation stage terms that occur in plain text as links are disambiguated so they can be linked to the appropriate Wikipedia article. This is done by balancing the commonness (i.e. prior probability) of a sense with its relatedness to the surrounding context. Link detection starts with building n-grams in the document, and retaining those whose probability exceeds a threshold. This threshold is intended to discard nonsense phrases and stop words. All the remaining phrases are disambiguated using the classifier from the link disambiguation step. The automatically extracted links and the known links from the Wikipedia are finally used as training instances for standard supervised classifiers, which are supported by various features like link probability, generality or location and spread.

Wikify services can be trained on any linked document collection. However, as most of the linked phrases in a Wikipedia article link to other Wikipedia articles in the same language, an additional cross-lingual linking step is needed. In this deliverable we use the obvious and straightforward approach of exploiting the existing language links provided in the Wikipedia articles.

2.5 Cross-lingual Explicit Semantic Analysis based approach

Explicit Semantic Analysis (ESA) is a recent prominent example of explicit models, and developed by Gabrilovich and Markovitch in 2007. Explicit models are concept based retrieval model, that allow an explicit representation of the meaning of documents based on concepts. In an explicit model externally defined concepts are given. The ESA method represents semantics of natural language texts using natural concepts, and is thus comprehensible to human users.

Given external defined concepts \( C = \{ c_1, c_2, ..., c_n \} \) classical monolingual Explicit Semantic Analysis takes a document \( x \) represented by a term vector as input and maps it to a concept vector. This concept vector space (in this case the English Wikipedia is that concept space) is spanned by a given collection of documents like \( D_c = \{ D_1, D_2, ..., D_n \} \) in language \( L_e \) such that each dimension corresponds to a document. For standard vector space, synonyms contribute nothing to document similarity. For more details on CLESA see D4.1.1
3 Cross-lingual Annotation Web Services

This section describes the technical implementation of the techniques introduced in the previous section.

3.1 Named Entity Annotation Service

This web service takes the output of linguistic processing in WP2 as input, adds the Wikipedia annotations by matching the names of the detected entities against the Wikipedia titles.

<table>
<thead>
<tr>
<th>Language</th>
<th>URL SandBox</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>English Service</td>
<td><a href="http://km.aifb.kit.edu/services/ner-annotation-en/">http://km.aifb.kit.edu/services/ner-annotation-en/</a></td>
<td><code>&lt;item&gt;</code>&lt;sentences&gt;&lt;sentence id=&quot;&quot;&gt;&lt;text&gt;&lt;token pos=&quot;&quot; end=&quot;&quot; lemma=&quot;&quot; id=&quot;&quot; start=&quot;&quot;&lt;/token&gt;&lt;/token&gt;&lt;/sentence&gt;&lt;/sentences&gt;&lt;entities&gt;&lt;entity type=&quot;&quot; displayName=&quot;&quot; id=&quot;&quot;&gt; &lt;mentions&gt;&lt;mention sentenceId=&quot;&quot; id=&quot;&quot; words=&quot;&quot; &lt;/mention&gt;&lt;/mentions&gt;&lt;/entity&gt;&lt;/entities&gt;</td>
</tr>
<tr>
<td>Spanish Service</td>
<td><a href="http://km.aifb.kit.edu/services/ner-annotation-es/">http://km.aifb.kit.edu/services/ner-annotation-es/</a></td>
<td>Same as English Service</td>
</tr>
<tr>
<td>Catalan Service</td>
<td><a href="http://km.aifb.kit.edu/services/ner-annotation-ca/">http://km.aifb.kit.edu/services/ner-annotation-ca/</a></td>
<td>Same as English Service</td>
</tr>
<tr>
<td>German Service</td>
<td><a href="http://km.aifb.kit.edu/services/ner-annotation-de/">http://km.aifb.kit.edu/services/ner-annotation-de/</a></td>
<td>Same as English Service</td>
</tr>
<tr>
<td>Chinese Service</td>
<td><a href="http://km.aifb.kit.edu/services/ner-annotation-zh/">http://km.aifb.kit.edu/services/ner-annotation-zh/</a></td>
<td>Same as English Service</td>
</tr>
<tr>
<td>Slovenian Service</td>
<td><a href="http://km.aifb.kit.edu/services/ner-annotation-sl/">http://km.aifb.kit.edu/services/ner-annotation-sl/</a></td>
<td>Same as English Service</td>
</tr>
</tbody>
</table>

Table 1. Named Entity Annotation Service
Input
<annotate>
<lang>de</lang>
<items>
    <sentence id="1">
        <text>Ich bin nach Zagreb gefahren</text>
        <tokens>
            <token id="1.1" pos="PPER" lemma="ich" start="0" end="3">Ich</token>
            <token id="1.2" pos="VAFIN" lemma="sein" start="4" end="7">bin</token>
            <token id="1.3" pos="APPR" lemma="nach" start="8" end="12">nach</token>
            <token id="1.4" pos="NE" lemma="Zagreb" start="13" end="19">Zagreb</token>
            <token id="1.5" pos="VPPP" lemma="fahren" start="20" end="28">fahren</token>
        </tokens>
    </sentence>
    <entities>
        <entity id="1" displayName="Zagreb" type="I-LOC">
            <mentions>
                <mention id="1.4" sentenceId="1" words="Zagreb" />
            </mentions>
        </entity>
    </entities>
</annotate>

Output
<item>
    <sentences>
        <sentence id="1">
            <text>Ich bin nach Zagreb gefahren</text>
            <tokens>
                <token id="1.1" pos="PPER" lemma="ich" start="0" end="3">Ich</token>
                <token id="1.2" pos="VAFIN" lemma="sein" start="4" end="7">bin</token>
                <token id="1.3" pos="APPR" lemma="nach" start="8" end="12">nach</token>
                <token id="1.4" pos="NE" lemma="Zagreb" start="13" end="19">Zagreb</token>
                <token id="1.5" pos="VPPP" lemma="fahren" start="20" end="28">fahren</token>
            </tokens>
        </sentence>
    </sentences>
    <entities>
        <entity id="1" displayName="Zagreb" type="I-LOC" concept="Zagreb">
            <mentions>
                <mention id="1.4" sentenceId="1" words="Zagreb" />
            </mentions>
        </entity>
    </entities>
    <annotations>
        <annotation displayName="Zagreb" lang="de" entityId="1">
            <mentions>
                <mention id="1.4" sentenceId="1" words="Zagreb" />
            </mentions>
            <descriptions>
                <description URL="http://de.wikipedia.org/wiki/Zagreb" lang="de" />
                <description URL="http://sl.wikipedia.org/wiki/Zagreb" lang="sl" />
                <description URL="http://zh.wikipedia.org/wiki/萨格勒布" lang="zh" />
                <description URL="http://ca.wikipedia.org/wiki/Zagreb" lang="ca" />
            </descriptions>
        </annotation>
    </annotations>
</item>

Figure 2. Example input and output of the named entity annotation service

3.2 Wikipedia Miner Wikifier Annotation Service

This web service takes the output of linguistic processing in WP2 as input, adds the Wikipedia annotations based on Wikifier service.
### 3.3 Cross-lingual Explicit Semantic Analysis Service

This web service is based on Explicit Semantic Analysis (ESA) and language links in Wikipedia. It uses Wikipedia dumps from May 2012 in English, German, Spanish, French, Catalan and Slovenian. The service can be called by using POST or GET request to the following URL address and input parameters.

**Service URL:** http://km.aifb.kit.edu/services/clesa/analyzer

**Input Parameters:**
- **doc:** the input document
- **lang1:** language of doc
- **lang2:** language of Wikipedia articles to retrieve
- **retrieve:** number of Wikipedia articles to retrieve

The response of the service consists of the xml elements **input** and **output**. The **input** element consists of a **doc** element, which contains the raw input document before pre-processing. The **output** element consists of a vector, which in turn contains the related concepts. The **concept** element has the following attributes:
- **lang:** the language of the related Wikipedia article
- **title:** the title of the related Wikipedia article
- **weight:** measure of the similarity between input document and the Wikipedia article

**Figure 3** shows the output format and returns the top ten related English Wikipedia articles.

---

**Table 2. Wikipedia Miner Wikifier Annotation Service**

<table>
<thead>
<tr>
<th>Language</th>
<th>URL Sandbox</th>
<th>Parameters</th>
</tr>
</thead>
</table>
| English Service| [http://km.aifb.kit.edu/services/annotation-en/](http://km.aifb.kit.edu/services/annotation-en/) | `<item>`
|                |                                   | `<sentences>`
|                |                                   | `<sentence id="">`
|                |                                   | `<text>
|                |                                   | `<tokens>`
|                |                                   | `<token pos=" " end=" " lemma=" " id=" " start="">
|                |                                   | `</token>`
|                |                                   | `</tokens>`
|                |                                   | `</sentence>`
|                |                                   | `</sentences>`
|                |                                   | `</entities>`
|                |                                   | `<entity type=" " displayName=" " id="">
|                |                                   | `<mentions>`
|                |                                   | `<mention sentenceId=" " id=" " words=""></mention>`
|                |                                   | `</mentions>`
|                |                                   | `</entity>`
|                |                                   | `</entities>`
|                |                                   | `</item>`
| Spanish Service| [http://km.aifb.kit.edu/services/annotation-es/](http://km.aifb.kit.edu/services/annotation-es/) | Same as English Service |
| German Service | [http://km.aifb.kit.edu/services/annotation-de/](http://km.aifb.kit.edu/services/annotation-de/) | Same as English Service |
The national executive of the strife-torn Democrats last night appointed little-known West Australian senator Brian Greig as interim leader - a shock move likely to provoke further conflict between the party’s senators and its organisation. In a move to reassert control over the party’s seven senators, the national executive last night rejected Aden Ridgeway’s bid to become interim leader, in favour of Senator Greig, a supporter of deposed leader Natasha Stott Despoja and an outspoken gay rights activist.

Figure 3. Example output of the cross-lingual ESA service
4 Evaluation of Cross-lingual Annotation

In this section, we present the evaluation results. In our evaluation the underlying assumption is, that the state-of-the-art approaches for monolingual annotation and grounding in Wikipedia perform best on English documents and the English Wikipedia. Source languages for evaluation in this deliverable were limited to German and Spanish only in order to obtain a proof of concept before extending the processing to other XLike languages. Thus, the evaluation is focused on annotating phrases in non-English documents (source language) and link them to English Wikipedia articles (target language).

At first we tried an automatic evaluation by comparing the found links with the ones found by the monolingual English annotation services. Since this processing was done on a parallel English-German-Spanish corpus (JRC-Acquis), it is expected that the same annotations should be found in any language which makes the detected links comparable. In this way we can calculate the precision of cross-lingual groundings of German and Spanish annotations in the English Wikipedia. We did this by counting the number of same links to the English Wikipedia found by both, the mono-lingual English annotation service and the cross-lingual source language (German/Spanish) annotation service and divide it by the total number of links extracted in the source language. There result showed a very low overlap (most of the time less then 1 out of 10) which we cannot explain yet. Thus, we decided not to speculate how to interpret this and do a manual evaluation. The inspection and interpretation of the automatic evaluation is postponed as future work.

Thus, the automatically inserted links to English Wikipedia were manually evaluated by marking the correctness of the links to English Wikipedia either as yes, no or 0, where yes and no were marking the correct or incorrect link respectively and 0 marked the link to Wikipedia disambiguation page. In processing of this evaluation results we took the conservative approach and treated 0 answers as no, so the calculated precision is representing the completely correct links (i.e, only links marked with yes).

While all implemented tools support many languages (NER: all XLike languages; WIFI: English, German, Spanish; CLESA: all XLike languages) for practical reasons and as a proof of concept, the evaluation presented here, is focused on annotating 88 German and Spanish documents only and linking them to the English Wikipedia. The English Wikipedia can be considered a hub knowledge base (but not a language independent knowledge base), as it is by far the largest and best linked Wikipedia, including language links.

4.1 Evaluation of Cross-lingual NER Annotation Service

First we counted the number of named entities detected in the source documents. Figure 4 shows the difference between the average number of extracted NEs per document by the NER service for English compared to German and Spanish. This confirms our assumption that the NER service for English performs differently than NER services for the other two languages and that our decision to use English as the hub language was appropriate.
Figure 4. Average number of NE links to English Wikipedia extracted per document

The precision of extracted annotations shown in Figure 5 also shows a differences for the different languages. The similarity of the German NER service to the English benchmark service seems to be much lower than to the Spanish one. Even the precision of the English service is somewhat below expectation. This has to be further investigated in the following development.

Figure 5. Average precision of NE links to English Wikipedia
4.2 Evaluation of Cross-lingual WIFI Annotation Service

Similar to the NER services there is a significant gap of the numbers of extracted annotation for the different languages. Again, in English documents more annotations are found compared to German and Spanish. This is probably due to the imbalance in the size (in terms of links and pages) of the English Wikipedia to the German and Spanish one.

Concerning the precision, as before, the German service detects more identical links found by the English service compared to the Spanish service, but for all three languages precision is above 92% which can be considered a very good result for the automatic mapping to the English Wikipedia. This precision will be checked for other XLike languages as well. Considering that a lot more links are extracted compared to the NER services (around 40 vs 7) this approach should also be significantly better in terms of recall.

Figure 6. Average number of Wikifier links to English Wikipedia extracted from one document

Figure 7. Average precision of Wikifier links to English Wikipedia
4.3 Evaluation of CLESA Annotation Service

In this set of experiments we investigate if the top Wikipedia topics (Wiki Topics, WT) associated to the English document is the same as the one associated to the source language (German and/or Spanish) and linked to the English language. Here the precision is lower than in the previous approach and the German system is clearly performing the lowest. These results will be checked for other XLike languages.

![Wiki Topics - Precision](image.png)

Figure 8. Average precision of CLESA links to English Wikipedia

4.4 Summary of Evaluation

As the summary of evaluation we calculated cumulative precision levels of possible combinations of these three approaches to cross-lingual linking to English Wikipedia articles. The averaged precisions over the different combination of methods is shown in Figure 9. Combination of Wikifier (WIKI) and Wiki Topics (WT) gives the combined precision around 90% and, if proved for other XLike languages, it can be a good ground for further research.
Our experiments show, that there is need for improvement to the naïve approach evaluated here of cross-lingual annotation by combining established techniques with Wikipedia language links. However, this still indicates, that a) the performance varies greatly between services trained for different languages and b) the average precision might need improvement to achieve even better performance.

A truly cross-lingual joint inference approach on a deeper semantic level is needed to level out the differences between languages and boost general precision. We intend to tackle this task in D3.1.2.
References

